Structure, Vol. 13, 857–860, June, 2005, ©2005 Elsevier Ltd All rights reserved. DOI 10.1016/j.str.2005.03.011

Structural Genomics of *Thermotoga maritima* Proteins Shows that Contact Order Is a Major Determinant of Protein Thermostability

Marc Robinson-Rechavi* and Adam Godzik
Joint Center for Structural Genomics
University of California, San Diego
9500 Gilman Drive
La Jolla, California 92093

Summary

Despite numerous studies, understanding the structural basis of protein stability in thermophilic organisms has remained elusive. One of the main reasons is the limited number of thermostable protein structures available for analysis, but also the difficulty in identifying relevant features to compare. Notably, an intuitive feeling of "compactness" of thermostable proteins has eluded quantification. With the unprecedented opportunity to assemble a data set for comparative analyses due to the recent advances in structural genomics, we can now revisit this issue and focus on experimentally determined structures of proteins from the hyperthermophilic bacterium *Thermotoga maritima*. We find that 73% of *T. maritima* proteins have higher contact order than their mesophilic homologs. Thus, contact order, a structural feature that was originally introduced to explain differences in folding rates of different protein families, is a significant parameter that can now be correlated with thermostability.

Introduction

Whereas most organisms do not survive at temperatures above 50°C, a small number of hyperthermophilic bacteria and archea thrive at temperatures over 80°C and in some cases greater than 110°C (Sterner and Liebl, 2001). This adaptation implies that molecular components of the cell, notably the proteins, should be stable at such high temperatures. Experimental studies of proteins from thermophilic organisms have demonstrated that they are usually stable and functional at high temperatures in vitro (Daniel and Danson, 2001). Understanding the molecular basis of thermostability is a very interesting and important research problem, both for our fundamental knowledge of protein structures and for the potential biotechnological applications of thermostable proteins, such as in bioremediation and high-temperature industrial processes.

Many structural features have been linked to protein thermostability, but the complexity of structural information and the heterogeneous sources of data have led to a confused picture (Petsko, 2001). One feature which seems intuitively to be related to stability is compactness of the protein, yet its characterization has proved elusive. Several statistical studies of protein structures from thermophiles have addressed this question using different proxies of "compactness." The simplest measure is protein length (longer proteins have larger and better packed buried volumes), but it differs little or not at all among homologous proteins from thermophiles and mesophiles (Das and Gerstein, 2000; our unpublished data). Various groups focused on compactness of thermostable proteins, with Kumar et al. (2000) using accessible surface area and Szilagyi and Zavodszky (2000) using cavity size as measured by total surface area of cavities, but none have found significant differences. Several of these studies report more residues in α helices and less residues in loops or disordered regions in proteins from thermophiles (Kumar et al., 2000; Szilagyi and Zavodszky, 2000; Chakravarty and Varadarajan, 2002), but this difference was not the main feature in any of these studies, and we were unable to confirm it independently (our unpublished data). Chakravarty and Varadarajan (2002) found some variation in solvent accessibility by distinguishing polar and nonpolar residues, and many groups identified an increase in the number of ion pairs inside thermostable proteins, which may point out to stronger packing in such proteins.

It seems then that either protein compactness is not such an important feature of proteins from thermophiles, or that an appropriate measure has not yet been found. In this study, we have investigated contact order (Plaxco et al., 1998), a structural parameter describing packing topology in proteins and found to correlate well with folding rates. We found a very significant trend: in 73% of the protein pairs analyzed, the thermostable homolog has a higher contact order. Apart from the choice of relevant parameters to compare, a drawback in such studies until recently was the paucity of data. To obtain experimental data sets for 25 pairs of structures (e.g., Szilagyi and Zavodszky, 2000), proteins from thermophilic archea and bacteria were grouped together and compared to a mixed set of bacterial, archeal, and sometimes eukaryotic homologs. To increase sample size, some authors made models of structures of proteins from thermophiles (e.g., Chakravarty and Varadarajan, 2002), which has the added risk that the modeling step may itself introduce bias in the observations. We have taken advantage of the recent and rapid increase in structural data from *Thermotoga maritima*, a hyperthermophilic bacteria with an optimum growth temperature of 80°C. Most of these new structures come from efforts at our PSI structural genomics center, the Joint Center for Structural Genomics (Lesley et al., 2002). For each *T. maritima* structure, we have identified a homolog with an experimentally determined structure and estimated the type of the homology relationship by phylogenetic analysis. Only orthologs or paralogs were retained, while laterally transferred genes were eliminated from further analysis. Moreover, we have limited our sampling of mesophiles to bacteria. Seventy-three pairs (47 orthologs and 26 paralogs) fulfilling our criteria were identified, with an average rmsd of 2.2 Å. This is the largest data set of this type studied so far and also the only one to our knowledge with well defined homology relationships between proteins.

*Correspondence: marc@sdsc.edu

Table 1. Output of Statistical Tests on Contact Order

Contact Threshold	Homology Relation	N	*T. maritima* Mean	Mesophilic Mean	Mean Difference	T Test	Wilcoxon Signed Ranks Test
6 Å	orthologs	47	0.0877	0.0842	0.0034	p = 0.0036	p = 0.001
	paralogs	26	0.0919	0.0868	0.0050	p = 0.0028	p = 0.007
	all	73	0.0892	0.0852	0.0040	p = 0.0003	p < 0.001
4.5 Å	orthologs	47	0.0630	0.0591	0.0038	p = 0.0003	p < 0.001
	paralogs	26	0.0655	0.0609	0.0046	p = 0.043	p = 0.007
	all	73	0.0639	0.0598	0.0041	$p < 10^{-4}$	p < 0.001

Results and Discussion

More than 70% of protein pairs we analyzed have higher relative contact order in *T. maritima* (53 out of 73). The average relative contact order of the *T. maritima* proteins is 0.0892, whereas the average of their close mesophilic homologs is 0.0852, a small but highly significant difference (paired t test, p = 0.0003; Table 1; and see Table S1 in the Supplemental Data available with this article online). The difference in contact order correlates to the solvent accessibility only in a subset of proteins (our unpublished data), and it is not related to any significant difference in secondary structure. The trend is not dependent on the type of homology relation: higher contact order is observed for orthologs as well as paralogs (Table 1). Such a significant difference is not recovered for distant homology pairs, detected by fold recognition, probably because major changes in function are dominant over any trends associated with thermostability at this scale (data not shown). We verified that these differences were not specific to *T. maritima* by comparing 31 pairs of close homologs between *T. maritima* and other thermophiles (18 bacterial and 13 archeal proteins); contact order is not significantly different in other thermophiles (0.0800 versus 0.0798; p = 0.95). Thus the difference we observed between *T. maritima* and mesophiles is related to the thermostability of *T. maritima* proteins. Of note, the increase in compactness is not due to any significant shortening of the loops in our data set (p = 0.47), in contradiction with previous reports (Thompson and Eisenberg, 1999).

Higher contact order has never been reported previously in association with thermostability, at least to our knowledge. Importantly, contact order is correlated to the folding rate of proteins that fold by a two-state kinetics (Plaxco et al., 1998; Makarov et al., 2002). This correlation has been confirmed by comparing homologs (Bemporad et al., 2004) or experimental mutants (Mason et al., 2002). Thus, the addition of one cysteine bridge can be enough to change significantly both contact order and folding rate (Mason et al., 2002), proving the possibility of significant differences between otherwise similar proteins, such as those we compared. Indeed, homologous proteins in our data set differ significantly in contact order despite very similar structures (average rmsd = 2.2 Å). There are specific cases with direct experimental evidence that a few key point "mutations" can change the flexibility, the thermostability, and the contact order in a manner consistent with our observations. For instance, a modified version of the artificial miniprotein BBAT1, "peptide 1" (PDB: 1sn9), was reported (Ali et al., 2004), which has higher thermostability (40°C to 64°C). The thermostable "peptide 1" has a higher contact order than BBA5, a variant of this protein that is not thermostable (PBD: 1t8j).

The highest difference in contact order in our data set is observed for GAPDH (glyceraldehyde 3-phosphate dehydrogenase), with an rmsd of only 1.18 Å between *T. maritima* and *E. coli* for a difference in contact order of 0.039 (Figure 1). Interestingly, this enzyme has been a highly studied example of thermostability (Korndorfer et al., 1995; Tanner et al., 1996; Song et al., 1998), yet a difference in contact order was not described previously, although support for higher compactness and rigidity was reported (Korndorfer et al., 1995). Over all bacterial GAPDHs, there appears to be a positive relation between contact order and growth temperature

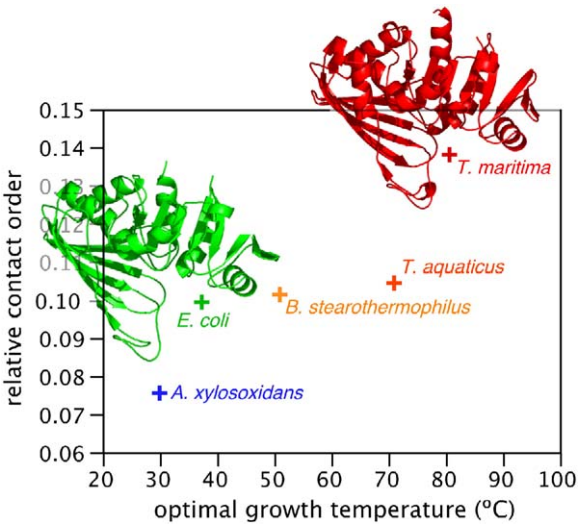

Figure 1. Comparison of GAPDH Structures from Bacteria

Relation between the optimal growth temperature and the relative contact order of GAPDH from different bacterial species. The x axis represents the mean optimal growth temperature for each species (Huang et al., 2004), and the y axis represents the relative contact order computed on one chain of the GAPDH homotetramer. Each point corresponds to one experimental structure from the PDB; points are color coded according to growth temperature. Chain O of the structures used from *E. coli* and *T. maritima* are represented. *E. coli* GAPDH-A is the mesophilic bacterial GAPDH with the lowest rmsd to *T. maritima* GAPDH; they are paralogs. No structure is available for mesophilic bacterial orthologs of *T. maritima* GAPDH (i.e., GAPDH-C), and *T. maritima* does not have GAPDH-A. The PDB entries used are: 1obf (*A. xylosoxidans*; 28°C–37°C), 1gad (*E. coli*; 37°C), 1gd1 (*B. stearothermophilus*; 37°C–65°C), 1cer (*T. aquaticus*; 70°C–72°C) and 1hdg (*T. maritima*; 80°C); in parenthesis, the range of optimal growth temperatures.

(Figure 1). GAPDHs from two thermophilic archea do not follow this trend, with contact orders of 0.074 and 0.077 for optimal growth temperatures of 87°C and 83°C; it has already been reported that archeal and bacterial GAPDHs have different thermostability strategies (Charron et al., 2002).

Galzitskaya et al. (2003) have shown that the correlation between contact order and folding rate holds only for proteins with two-state folding kinetics, but not for proteins with three-state folding kinetics; instead, folding rate in the latter is inversely correlated to protein length. Interestingly, the 20 proteins that have a decreased contact order in *T. maritima* also have greater average length (232 versus 224 amino acids). This is in contrast to generally shorter chains in *T. maritima* proteins than in mesophiles, including those with increased contact order. The difference in length variation between pairs with increased or decreased contact order is significant (Wilcoxon test: $p = 0.0038$). This suggests that lower folding rates in *T. maritima* than in mesophiles may be ubiquitous, through higher contact order for some proteins and through increased chain length in others, although this remains quite hypothetical. Of note, some experimental studies suggest similar folding rates, but lower unfolding rates, of proteins from thermophiles (Hollien and Marqusee, 2002). And as a note of caution, it should be kept in mind that the correlation between contact order and folding rate was established at room temperature (20°C–37°C; Table 1 in Plaxco et al. [1998]), not at 80°C.

The original definition of contact order, used in these calculations, sets a limit of 6 Å between atoms to define them as being "in contact" (Plaxco et al., 1998). This definition was proposed as a proxy for topological complexity, and the resulting contact order parameter has the obvious merit of being correlated to experimental measures of protein folding rates. But we wanted to test a more stringent definition of contact order, limited to those residues that are effectively in physical contact. Thus, we set a limit of 4.5 Å, which was previously established as the optimal value for the development of empirical interaction parameters in proteins (Godzik et al., 1992, 1995). This modified contact order definition varies even more significantly than the original (Table 1).

Interestingly, *T. maritima* proteins from this study also show a systematic and statistically significant difference in energy-like score calculated using the contact-based empirical energy parameters (Godzik et al., 1992, 1995). On the set of pairs analyzed here, in 52 out of 73 pairs the *Thermotoga* structure had a better score, with an average difference of –0.044 kT ($p \leq 0.0001$). Empirical scores like this are typically used to validate models in comparative modeling, and their correlation to experimentally measured protein stability was discussed in the literature, but never convincingly proven.

In conclusion, contact order is a major structural determinant of protein thermostability. It allows us to demonstrate that a specific structural feature distinguishes proteins from a thermophile from their mesophilic homologs. This feature is related to compactness of the structure, but also to its topological complexity. What is even more important, it indirectly relates thermostability to the protein folding/unfolding rate. Further functional and structural implications of this finding should be explored both experimentally and by bioinformatic analyses.

Experimental Procedures

All sequences associated with experimentally determined protein structures were downloaded from the PDB database (Bourne et al., 2004) version from November 1, 2004. The subset of entries from *Thermotoga maritima* was compared by BlastP (Altschul et al., 1997) to all other sequences. For each *T. maritima* entry which had at least one hit with an E value under e–4, aligned homologous proteins from completely sequenced genomes were recovered from Hogenom (Perrière et al., 2003). These alignments were edited to add sequences of homologous PDB entries that are absent from Hogenom (typically from organisms whose genome is not sequenced) and to merge protein families that were classified separately in Hogenom but were homologous according to the results of Blast on the PDB sequence set.

For each alignment a phylogenetic tree was built by PhyML (Guindon and Gascuel, 2003), with the JTT model and a γ distribution between sites (parameter α estimated by PhyML with eight categories). These trees were used to assess homology relations between the proteins from the PDB entries: orthology, paralogy, or xenology (lateral gene transfer). Five cases of suspected lateral transfer were excluded from the final data set. When several homologous proteins were available from *T. maritima*, only one was used.

Structure coordinates were obtained from the PDB (Bourne et al., 2004), and only single chains were used in further analysis. For homo-oligomers, only the first chain in the file was used in calculations. For hetero-oligomers, chains were treated separately according to their homology relations as established by phylogenetic analysis. When there were several mesophilic bacterial proteins orthologous to a same *T. maritima* proteins, the one with the lowest rmsd was used. The same was done for paralogous proteins. When structures of both orthologous and paralogous proteins were available, orthologs were preferred.

Relative contact order was calculated by the program contactOrder.pl (http://depts.washington.edu/bakerpg/contact_order/), which implements the definition of Plaxco et al. (1998): any nonwater atoms separated by less than 6 Å are considered "in contact." In addition, we calculated contact order defining contact with a more stringent distance of 4.5 Å, which optimizes the detection of specific interactions (Godzik et al., 1992, 1995).

Supplemental Data

Supplemental Data can be found online at http://www.structure.org/cgi/content/full/13/6/857/DC1/.

Acknowledgments

We thank Iddo Friedberg, Olga Kirillova and Ian A. Wilson for helpful suggestions, as well as all members of the Joint Center for Structural Genomics.

Received: December 20, 2004
Revised: March 17, 2005
Accepted: March 17, 2005
Published: June 7, 2005

References

Ali, M.H., Peisach, E., Allen, K.N., and Imperiali, B. (2004). X-ray structure analysis of a designed oligomeric miniprotein reveals a discrete quaternary architecture. Proc. Natl. Acad. Sci. USA *101*, 12183–12188.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST:

a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Bemporad, F., Capanni, C., Calamai, M., Tutino, M.L., Stefani, M., and Chiti, F. (2004). Studying the folding process of the acylphosphatase from *Sulfolobus solfataricus*. A comparative analysis with other proteins from the same superfamily. Biochemistry *43*, 9116–9126.

Bourne, P.E., Addess, K.J., Bluhm, W.F., Chen, L., Deshpande, N., Feng, Z., Fleri, W., Green, R., Merino-Ott, J.C., Townsend-Merino, W., et al. (2004). The distribution and query systems of the RCSB Protein Data Bank. Nucleic Acids Res. *32*, D223–D225.

Chakravarty, S., and Varadarajan, R. (2002). Elucidation of factors responsible for enhanced thermal stability of proteins: a structural genomics based study. Biochemistry *41*, 8152–8161.

Charron, C., Vitoux, B., and Aubry, A. (2002). Comparative analysis of thermoadaptation within the archaeal glyceraldehyde-3-phosphate dehydrogenases from mesophilic *Methanobacterium bryantii* and thermophilic *Methanothermus fervidus*. Biopolymers *65*, 263–273.

Daniel, R.M., and Danson, M.J. (2001). Assaying activity and assessing thermostability of hyperthermophilic enzymes. Methods Enzymol. *334*, 283–293.

Das, R., and Gerstein, M. (2000). The stability of thermophilic proteins: a study based on comprehensive genome comparison. Funct. Integr. Genomics *1*, 76–88.

Galzitskaya, O.V., Garbuzynskiy, S.O., Ivankov, D.N., and Finkelstein, A.V. (2003). Chain length is the main determinant of the folding rate for proteins with three-state folding kinetics. Proteins *51*, 162–166.

Godzik, A., Kolinski, A., and Skolnick, J. (1992). Topology fingerprint approach to the inverse protein folding problem. J. Mol. Biol. *227*, 227–238.

Godzik, A., Kolinski, A., and Skolnick, J. (1995). Are proteins ideal mixtures of amino acids? Analysis of energy parameter sets. Protein Sci. *4*, 2107–2117.

Guindon, S., and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. *52*, 696–704.

Hollien, J., and Marqusee, S. (2002). Comparison of the folding processes of *T. thermophilus* and *E. coli* Ribonucleases H1. J. Mol. Biol. *316*, 327–340.

Huang, S.-L., Wu, L.-C., Liang, H.-K., Pan, K.-T., Horng, J.-T., and Ko, M.-T. (2004). PGTdb: a database providing growth temperatures of prokaryotes. Bioinformatics *20*, 276–278.

Korndorfer, I., Steipe, B.S., Huber, R., Tomschy, A., and Jaenicke, R. (1995). The crystal structure of holo-glyceraldehyde-3-phosphate dehydrogenase from the hyperthermophilic bacterium *Thermotoga maritima* at 2.5 A resolution. J. Mol. Biol. *246*, 511–521.

Kumar, S., Tsai, C.J., and Nussinov, R. (2000). Factors enhancing protein thermostability. Protein Eng. *13*, 179–191.

Lesley, S.A., Kuhn, P., Godzik, A., Deacon, A.M., Mathews, I., Kreusch, A., Spraggon, G., Klock, H.E., McMullan, D., Shin, T., et al. (2002). Structural genomics of the *Thermotoga maritima* proteome implemented in a high-throughput structure determination pipeline. Proc. Natl. Acad. Sci. USA *99*, 11664–11669.

Makarov, D.E., Keller, C.A., Plaxco, K.W., and Metiu, H. (2002). How the folding rate constant of simple, single-domain proteins depends on the number of native contacts. Proc. Natl. Acad. Sci. USA *99*, 3535–3539.

Mason, J.M., Gibbs, N., Sessions, R.B., and Clarke, A.R. (2002). The influence of intramolecular bridges on the dynamics of a protein folding reaction. Biochemistry *41*, 12093–12099.

Perrière, G., Combet, C., Penel, S., Blanchet, C., Thioulouse, J., Geourjon, C., Grassot, J., Charavay, C., Gouy, M., Duret, L., and Deléage, G. (2003). Integrated databanks access and sequence/structure analysis services at the PBIL. Nucleic Acids Res. *31*, 3393–3399.

Petsko, G.A. (2001). Structural basis of thermostability in hyperthermophilic proteins, or "There's more than one way to skin a cat". Methods Enzymol. *334*, 469–478.

Plaxco, K.W., Simons, K.T., and Baker, D. (1998). Contact order, transition state placement and the refolding rates of single domain proteins. J. Mol. Biol. *277*, 985–994.

Song, S.Y., Li, J., and Lin, Z.J. (1998). Structure of holo-glyceraldehyde-3-phosphate dehydrogenase from *Palinurus versicolor* refined at 2 angstrom resolution. Acta Crystallogr. D Biol. Crystallogr. *54*, 558–569.

Sterner, R., and Liebl, W. (2001). Thermophilic adaptation of proteins. Crit. Rev. Biochem. Mol. Biol. *36*, 39–106.

Szilagyi, A., and Zavodszky, P. (2000). Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey. Struct. Fold. Des. *8*, 493–504.

Tanner, J.J., Hecht, R.M., and Krause, K.L. (1996). Determinants of enzyme thermostability observed in the molecular structure of *Thermus aquaticus* D-glyceraldehyde-3-phosphate dehydrogenase at 2.5 angstrom resolution. Biochemistry *35*, 2597–2609.

Thompson, M.J., and Eisenberg, D. (1999). Transproteomic evidence of a loop-deletion mechanism for enhancing protein thermostability. J. Mol. Biol. *290*, 595–604.